

Rademacher Complexity of the Restricted Boltzmann Machine

ASYMPTOTIC CONDITION AND CD-1 APPROXIMATION

December 8, 2015

Xiao (Cosmo) Zhang

Department of Computer Science

zhang923@purdue.edu

Abstract

Boltzmann machine, as a fundamental construction block of deep belief network and deep Boltzmann machines, is widely used in deep learning community and great success has been achieved. However, theoretical understanding of many aspects of it is still far from clear. In this paper, we studied the Rademacher complexity of both the asymptotic restricted Boltzmann machine and the practical implementation with single-step contrastive divergence (CD-1) procedure. Our results disclose the fact that practical implementation training procedure indeed increased the Rademacher complexity of restricted Boltzmann machines. A further research direction might be the investigation of the VC dimension of a compositional function used in the CD-1 procedure.

1 Introduction

A restricted Boltzmann machine (RBM) is a generative graphical model that can learn a probability distribution over its set of inputs. Initially proposed by Smolensky [1986] for modeling cognitive process, it grew to prominence after successful application were found by Geoffrey Hinton and his collaborators [Hinton and Salakhutdinov, 2006, 2012; Salakhutdinov and Hinton, 2009]. As a building block for deep belief network (DBN) and deep Boltzmann machines (DBM), RBM is extremely useful for pre-training the data by projecting them to a hidden layer. Also, it is proved that by adding another layer on top of a RBM, the variational lower bound of the data likelihood can be increased [Hinton et al., 2006; Salakhutdinov and Hinton, 2012], which conveys the theoretical advantage of building multilayer RBMs.

Pre-training of the data by using a RBM is essentially a unsupervised learning process, in which no label of the data is provided. Instead, the training process is trying to maximize the data likelihood by finding a proper set of parameters of the RBM.

However less attention has been given to the analysis of Rademacher complexity on RBMs. Rademacher complexity in the computational learning theory, measures richness of a class of real-valued functions with respect to a probability distribution. It can be regarded as a generalization of PAC-Bayes analysis. Its particular setting can help analysis of unsupervised learning algorithms,

rather than merely the prediction problems, given the hypothesis class is possibly infinite. Honorio [2012] also proved that discrete factor graphs, including Markov random fields, are Lipschitz continuous, which motivated this work to further investigate the properties of RBM.

The goal of this paper is trying to bound the Rademacher complexity for the likelihood of the RBM algorithm from a given training data set, with pre-assumptions that the model structure of the RBM is known (data dimensionality and number of hidden nodes).

2 Preliminaries

In the beginning of this section we introduce Lipschitz continuity.

Definition 1. *Given the parameters $\Theta \in \mathbb{R}^{M_1 \times M_2}$, a differentiable function $f(\Theta) \in \mathbb{R}$ is called Lipschitz continuous with respect to the l_p -norm of Θ , if there exist a constant $K \geq 0$ such that:*

$$(\forall \Theta_1, \Theta_2) |f(\Theta_1) - f(\Theta_2)| \leq K \|\Theta_1 - \Theta_2\| \quad (1)$$

or equivalently:

$$(\forall \Theta) \left\| \frac{\partial f}{\partial \Theta} \right\| \leq K \quad (2)$$

Next we introduce the Rademacher complexity.

Definition 2.

Definition 2.1. *A random variable $x \in \{-1, +1\}$ is called Rademacher $\iff \mathbb{P}(x) \sim \text{Bernoulli}(0.5)$.*

Definition 2.2. *The empirical Rademacher complexity of the hypothesis class \mathcal{H} w.r.t a data set $\mathcal{S} = \{z^{(1)} \dots z^{(n)}\}$ is defined as:*

$$\hat{\mathfrak{R}}_s(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i h(z^{(i)}) \right) \right] \quad (3)$$

At the end of this section we give formal definition of the restricted Boltzmann machine.

Definition 3. *The restricted Boltzmann machine is a two layer Markov Random Field, where the observed binary stochastic visible units $\mathbf{x} \in \{0, 1\}^k$ have pairwise connections to the binary stochastic hidden units $\mathbf{h} \in \{0, 1\}^m$. There are no pairwise connections within the visible units, nor within the hidden ones. Restricted Boltzmann machine is a energy-based model, in which we define the energy for a state $\{\mathbf{x}, \mathbf{h}\}$ as*

$$\text{Energy}(\mathbf{x}, \mathbf{h}; \theta) = -\mathbf{x}^T \mathbf{b} - \mathbf{h}^T \mathbf{c} - \mathbf{x}^T \mathbf{W} \mathbf{h}, \quad (4)$$

where $\theta = \{\mathbf{c}, \mathbf{b}, \mathbf{W}\}$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^k$, and $\mathbf{W} \in \mathbb{R}^{k \times m}$. Hence, we can write the likelihood for an observation \mathbf{x} as

$$p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} \exp\{-Energy(\mathbf{x}, \mathbf{h}; \theta)\}}{Z_{\theta}}, \quad (5)$$

where

$$Z_{\theta} = \sum_{\mathbf{h}} \sum_{\mathbf{x}} \exp\{-Energy(\mathbf{x}, \mathbf{h}; \theta)\}, \quad (6)$$

which is the partition function, used for normalization. The sum over \mathbf{h} and \mathbf{x} enumerate all the possible values for the visible units and the hidden ones. Our goal optimization is to maximize the log-likelihood (minimize the negative log-likelihood) of the model. For N data samples, we can write the log-likelihood as:

$$\ln p_{\theta}(\mathbf{x}) = \underbrace{\ln\left\{\sum_{\mathbf{h}} \exp\{-Energy(\mathbf{x}, \mathbf{h}; \theta)\}\right\}}_1 - \underbrace{\ln Z_{\theta}}_2 \quad (7)$$

3 Rademacher Complexity

In this section, we provide a up-bound of the empirical Rademacher complexity for the likelihood of the restricted Boltzmann machine. Since part 2, the partition function of the restricted Boltzmann machine, of equation 7 is not depending on the data set. This part does not have any randomness and the Rademacher complexity of it is 0 by the definition. Thus, we can only focus on the Rademacher complexity of part 1 of equation 7. Denote \mathbf{W}_j as the j -th column of the matrix \mathbf{W} , c_j as the j -th element of \mathbf{c} , h_j as the j -th element of \mathbf{h} . By expanding part 1 of equation 7, we get

$$part\ 1 = \ln\left\{\sum_{\mathbf{h}} \exp\{-Energy(\mathbf{x}, \mathbf{h}; \theta)\}\right\} \quad (8)$$

$$= \ln\left\{\sum_{h_1} \cdots \sum_{h_m} \exp\left[\mathbf{x}^T \mathbf{b} + \sum_{j=1}^m \mathbf{x}^T \mathbf{W}_j h_j + \sum_{j=1}^m h_j c_j\right]\right\} \quad (9)$$

$$= \ln\left\{\prod_{j=1}^m \left[\sum_{h_j \in \{0,1\}} \exp\left(\mathbf{x}^T \mathbf{b} + \sum_{j=1}^m \mathbf{x}^T \mathbf{W}_j h_j + \sum_{j=1}^m h_j c_j\right)\right]\right\} \quad (10)$$

$$= \sum_{j=1}^m \ln [\exp(\mathbf{x}^T \mathbf{b}) + \exp(\mathbf{x}^T \mathbf{b} + \mathbf{x}^T \mathbf{W}_j + c_j)] \quad (11)$$

Lemma 1. Let $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^d\}$. Let \mathcal{F} be the class of linear predictors, i.e.,

$$\mathcal{F} = \{\mathbf{b}^T \mathbf{x} | \mathbf{b} \in \mathbb{R}^d \text{ and } \|\mathbf{b}\|_1 \leq B\}. \quad (12)$$

We have

$$\hat{\mathfrak{R}}_s(\mathcal{F}) \leq B \sqrt{\frac{2 \ln(d)}{n}} \quad (13)$$

Proof. Let $\mathcal{S} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ be a data set of n samples. Denote x_j as the j -th element of \mathbf{x} .

$$\hat{\mathfrak{R}}_s(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{x}^{(i)}) \right) \right] \quad (14)$$

$$= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^n \sigma_i \mathbf{b}^T \mathbf{x}^{(i)} \right) \right] \quad (15)$$

$$= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\mathbf{b}: \|\mathbf{b}\|_1 \leq B} \left(\mathbf{b}^T \left(\sum_{i=1}^n \sigma_i \mathbf{x}^{(i)} \right) \right) \right] \quad (16)$$

$$= \frac{B}{n} \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{x}^{(i)} \right\|_\infty \right] \quad (17)$$

$$= \frac{B}{n} \mathbb{E}_\sigma \left[\sup_{j \in \{1 \dots d\}} \left(\sum_{i=1}^n \sigma_i \mathbf{x}^{(i)} \right)_j \right] \quad (18)$$

$$\leq \frac{B \sqrt{2 \ln(d)}}{n} \sup_{j \in \{1 \dots d\}} \sqrt{\sum_{i=1}^n \left[x_j^{(i)} \right]^2} \quad (19)$$

$$\leq \frac{B \sqrt{2 \ln(d)}}{n} \sqrt{n \|\mathbf{x}\|_\infty^2} \quad (20)$$

$$= \|\mathbf{x}\|_\infty B \sqrt{\frac{2 \ln(d)}{n}} \quad (21)$$

$$= B \sqrt{\frac{2 \ln(d)}{n}} \quad (22)$$

Equation 17 uses Holder's inequality when the equal sign is taken. inequality 19 uses Massart's finite class lemma. Equation 22 is from the fact that $\mathbf{x} \in \{0, 1\}^d$. Therefore we proved inequality 13. \square

Remark 1. *Function*

$$\phi(g) = \ln(1 + \exp(g)) \quad (23)$$

is 1-Lipschitz continuous for $g \in \mathbb{R}$.

Proof. $|\partial \phi(g) / \partial g| = \text{Sigmoid}(g) \leq 1$. \square

Lemma 2. Let $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^d\}$, \mathcal{F} be a class of linear predictors, i.e.,

$$\mathcal{F} = \{\mathbf{b}^T \mathbf{x} | \mathbf{b} \in \mathbb{R}^d\}. \quad (24)$$

Let \mathcal{G} be another class of linear predictors, i.e.,

$$\mathcal{G} = \{\mathbf{w}^T \mathbf{x} + c | \mathbf{w} \in \mathbb{R}^d, c \in \mathbb{R}\}. \quad (25)$$

Let \mathcal{H} be a function of \mathcal{F} and \mathcal{G} , written as

$$\mathcal{H} = \{\ln[\exp(f(\mathbf{x})) + \exp(f(\mathbf{x}) + g(\mathbf{x}))] | f \in \mathcal{F}, g \in \mathcal{G}\}, \quad (26)$$

Let $\mathcal{S} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ be a data set of n samples. We have

$$\hat{\mathfrak{R}}_s(\mathcal{H}) \leq \hat{\mathfrak{R}}_s(\mathcal{F}) + \hat{\mathfrak{R}}_s(\mathcal{G}) \quad (27)$$

Proof.

$$\hat{\mathfrak{R}}_s(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i h(\mathbf{x}^{(i)}) \right) \right] \quad (28)$$

$$= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \ln [\exp(f(\mathbf{x}^i)) + \exp(f(\mathbf{x}^i) + g(\mathbf{x}^i))] \right) \right] \quad (29)$$

$$= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \ln [\exp(f(\mathbf{x}^i))] + \ln [1 + \exp(g(\mathbf{x}^i))] \right) \right] \quad (30)$$

$$\leq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{x}^i) \right) \right] + \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \ln [1 + \exp(g(\mathbf{x}^i))] \right) \right] \quad (31)$$

$$= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{x}^i) \right) \right] + \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \phi(g(\mathbf{x}^i)) \right) \right] \quad (32)$$

$$\leq \hat{\mathfrak{R}}_s(\mathcal{F}) + \hat{\mathfrak{R}}_s(\mathcal{G}) \quad (33)$$

□

In inequality 31, the first part of it is exactly the Rademacher complexity of \mathcal{F} by definition. The second part can be shown to be $\leq \hat{\mathfrak{R}}_s(\mathcal{G})$ by using Ledoux-Talagrand Contraction Lemma, combining with the results in **Remark 1** that $\phi(g)$ is 1-Lipschitz continuous. Hence we proved inequality 27.

Remark 2. Let $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^d\}$. Let \mathcal{G} be the class of linear predictors, i.e.,

$$\mathcal{G} = \{\mathbf{w}^\top \mathbf{x} + c | \mathbf{w} \in \mathbb{R}^d, c \in \mathbb{R} \text{ and } \|\mathbf{w}\|_1 \leq W\}, \quad (34)$$

where \mathbf{W}_j is the j -th column of \mathbf{W} . Let $\mathcal{S} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ be a data set of n samples. We have

$$\hat{\mathfrak{R}}_s(\mathcal{G}) \leq W \sqrt{\frac{2 \ln(d)}{n}} \quad (35)$$

Proof.

$$\hat{\mathfrak{R}}_s(\mathcal{G}) = \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g(\mathbf{x}^{(i)}) \right) \right] \quad (36)$$

$$= \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbf{w}^\top \mathbf{x}^{(i)} + c) \right) \right] \quad (37)$$

$$\leq \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq W} \left(\mathbf{w}^\top \left(\sum_{i=1}^n \sigma_i \mathbf{x}^{(i)} \right) \right) + \sup_c \left(\sum_{i=1}^n \sigma_i \right) \right] \quad (38)$$

$$= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq W} \left(\mathbf{w}^\top \left(\sum_{i=1}^n \sigma_i \mathbf{x}^{(i)} \right) \right) \right] + \frac{1}{n} \mathbb{E}_\sigma \left[\sup_c \left(\sum_{i=1}^n \sigma_i \right) \right] \quad (39)$$

$$(40)$$

Notice that the first part of equation 39 can be bounded by $W \sqrt{\frac{2 \ln(d)}{n}}$ by using the results in **Lemma 1**, and the second part is exactly 0 by the definition of Rademacher complexity. Thus we proved inequality 35. \square

Theorem 1. Let $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^k\}$, $\mathcal{S} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ be a data set of n samples. Given a restricted Boltzmann machine with k visible units and m hidden ones. For all the parameters $\boldsymbol{\theta} = \{\mathbf{c}, \mathbf{b}, \mathbf{W}\}$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^k$, and $\mathbf{W} \in \mathbb{R}^{k \times m}$, assuming \mathbf{b}, \mathbf{W} are bounded by spheres $\|\mathbf{b}\|_1 \leq B$, $\|\mathbf{W}\|_{\max} = \forall j \|\mathbf{W}_j\|_1 \leq W$, where \mathbf{W}_j is the j -th column of \mathbf{W} . We can bound the empirical Rademacher complexity for the likelihood of this restricted Boltzmann machine as:

$$\hat{\mathfrak{R}}_s(\ln p_{\boldsymbol{\theta}}) \leq m \sqrt{\frac{2 \ln(k)}{n}} (B + W). \quad (41)$$

Proof. As we stated, we only consider equation 11 that has randomness and ignore the partition part of the log-likelihood. Using the notation in **Lemma 2**, we can write

$$\hat{\mathfrak{R}}_s(\ln p_{\boldsymbol{\theta}}) = \hat{\mathfrak{R}}_s\left(\sum_{j=1}^m \mathcal{H}_j\right) \leq \sum_{j=1}^m \hat{\mathfrak{R}}_s(\mathcal{H}_j), \quad (42)$$

which is from the elementary properties of the Rademacher complexity. Knowing the fact that each \mathcal{H}_j is of the same hypothesis space further provides us with

$$\sum_{j=1}^m \hat{\mathfrak{R}}_s(\mathcal{H}_j) = m \hat{\mathfrak{R}}_s(\mathcal{H}). \quad (43)$$

From **Lemma 2** we have $\hat{\mathfrak{R}}_s(\mathcal{H}) \leq \hat{\mathfrak{R}}_s(\mathcal{F}) + \hat{\mathfrak{R}}_s(\mathcal{G})$. And from **Lemma 1** we can directly bound $\hat{\mathfrak{R}}_s(\mathcal{F})$ by $B \sqrt{\frac{2 \ln(k)}{n}}$. For $\hat{\mathfrak{R}}_s(\mathcal{G})$, by using the results in **Remark 2**, we can bound it by $W \sqrt{\frac{2 \ln(k)}{n}}$. Together with inequality 42 and equation 43, we proved this theorem. \square

4 Rademacher Complexity with CD-1 Approximation

Contrastive Divergence is an approximation of the log-likelihood gradient that has been found to be a successful update rule for training RBMs. The reason that we are applying contrastive divergence algorithm is that because the partition function is can be hardly estimated by enumerating all the possible values because the complexity will be in the order of exponential, nor the factorization trick we used for the numerator can be used. In order to approximate the partition function for all possible visible examples, a MCMC chain is created. First an example is sampled uniformly from the empirical training examples. Then a mean-field approximation is applied to obtain the values of hidden units (whose values are also binary): Rather than sample from the distribution of \mathbf{h} , we use the values $\forall i, P(h_i = 1)$ as the values to approximate the samples. After we obtain $\tilde{\mathbf{h}}$, we have the distribution of \mathbf{x} based on the current values of \mathbf{h} (mean-field approximation) and parameters $(\mathbf{W}, \mathbf{b}, \mathbf{c})$. We sample from this distribution to obtain a vector $\tilde{\mathbf{x}}$ and use it to approximate the partition function. This procedure can also extended to more steps (CD-k, k steps). But experiments have shown that, even one step (CD-1) can yield a good performance for the model [Bengio, 2009].

After using CD-1 algorithm, the Rademacher complexity of the second part of equation 7 is no more free of randomness, due to the fact that $\tilde{\mathbf{x}}$ is a function of \mathbf{x} . If we rewrite the second part as

$$Z_{\theta} \approx \sum_{\mathbf{h}} \exp\{-Energy(\tilde{\mathbf{x}}, \mathbf{h}; \theta)\}, \quad (44)$$

Rademacher complexity of this term is also depending on random variable \mathbf{x} .

To simplify the procedure but without losing generality, instead of sampling $\tilde{\mathbf{x}}$ from its distribution, we also use mean field approximation to obtain $\tilde{\mathbf{x}}$. Also, we can write the energy function as

$$Energy(\mathbf{x}, \mathbf{h}; \theta) = \mathbf{x}^T \mathbf{W} \mathbf{h}, \quad (45)$$

while ignore the bias term for simplicity.

Remark 3. Using mean field approximation, we can obtain

$$\tilde{\mathbf{h}}^T = (sgm(\mathbf{x}^T \mathbf{W}_{\cdot 1}), \dots, sgm(\mathbf{x}^T \mathbf{W}_{\cdot m})), \quad (46)$$

where $sgm()$ is the sigmoid function, and $\mathbf{W}_{\cdot j}$ is the j -th column of \mathbf{W} .

$$\text{Proof. } P(\tilde{\mathbf{h}}|_{\mathbf{x}} = \mathbf{1}) = \frac{\exp\{\mathbf{x}^T \mathbf{W} \mathbf{1}\}}{\sum_{\mathbf{h}} \exp\{\mathbf{x}^T \mathbf{W} \mathbf{h}\}} = \frac{\prod_{i=1}^m \exp\{\mathbf{x}^T \mathbf{W}_{\cdot i}\}}{\prod_{i=1}^m \sum_{h_i} \exp\{\mathbf{x}^T \mathbf{W}_{\cdot i} h_i\}}$$

$$\text{With the fact that } \forall i, j \ h_i \perp h_j|_{\mathbf{x}}, \forall i, P(\tilde{h}_i|_{\mathbf{x}} = 1) = \frac{\exp\{\mathbf{x}^T \mathbf{W}_{\cdot i}\}}{1 + \exp\{\mathbf{x}^T \mathbf{W}_{\cdot i}\}} = sgm(\mathbf{x}^T \mathbf{W}_{\cdot i}). \quad \square$$

Remark 4. Similar to **Remark 3**, we can obtain

$$\tilde{\mathbf{x}}^T = (sgm(\mathbf{W}_{1 \cdot} \tilde{\mathbf{h}}), \dots, sgm(\mathbf{W}_{k \cdot} \tilde{\mathbf{h}})), \quad (47)$$

where $\forall v \ \mathbf{W}_{v \cdot}$ is the v -th row of \mathbf{W} .

Lemma 3. By using CD-1 Algorithm, and mean field approximation for both $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{h}}$, we have part 2 of equation 7 as

$$\ln Z_\theta = \sum_{j=1}^m \ln \left[1 + \exp \left\{ \sum_{i=1}^k \mathbf{W}_{ij} \text{sgm} \left(\sum_{v=1}^m \mathbf{W}_{iv} \text{sgm}(\mathbf{x}^\top \mathbf{W}_{\cdot v}) \right) \right\} \right], \quad (48)$$

where we use \mathbf{W}_{ij} to denote the element of i -th row and j -th column of matrix \mathbf{W} .

Proof. (sketch) Using the results from **Remark 3** and **Remark 4**, and the same factorization trick used before in equation 10, this can be shown easily. \square

Lemma 4. Let $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^k\}$, Let \mathcal{T} be a compositional function of \mathbf{x} with parameters \mathbf{W} , i.e.,

$$\mathcal{T} = \left\{ \mathbf{W}_{uj} \text{sgm} \left(\sum_{v=1}^m \mathbf{W}_{uv} \text{sgm}(\mathbf{x}^\top \mathbf{W}_{\cdot v}) \right) \mid \mathbf{W} \in \mathbb{R}^{k \times m}, \forall u \in \{1, \dots, k\}, \forall j \in \{1, \dots, m\} \right\}, \quad (49)$$

and assuming \mathbf{W} is bounded by spheres $\|\mathbf{W}\|_{\max} = \forall j \|\mathbf{W}_{\cdot j}\|_1 \leq W$, where $\mathbf{W}_{\cdot j}$ is the j -th column of \mathbf{W} . Also Let $\mathcal{S} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ be a data set of n samples. We have

$$\hat{\mathfrak{R}}_s(\mathcal{T}) \leq \frac{W \sqrt{2n \ln |\mathcal{T}|}}{n} \quad (50)$$

Proof.

$$\hat{\mathfrak{R}}_s(\mathcal{T}) = \mathbb{E}_\sigma \left[\sup_{t_{\mathbf{w}} \in \mathcal{T}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right) \right] \implies \quad (51)$$

$$\exp\{\mathbb{E}_\sigma \left[s \sup_{t_{\mathbf{w}} \in \mathcal{T}} \left(\sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right) \right]\} \leq \mathbb{E}_\sigma \left[\exp\{s \sup_{t_{\mathbf{w}} \in \mathcal{T}} \left(\sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right)\} \right] \quad (52)$$

$$= \sup_{t_{\mathbf{w}} \in \mathcal{T}} \mathbb{E}_\sigma \left[\exp\{s \left(\sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right)\} \right] \quad (53)$$

$$\leq \sum_{t_{\mathbf{w}} \in \mathcal{T}} \mathbb{E}_\sigma \left[\exp\{s \left(\sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right)\} \right] \quad (54)$$

$$= \sum_{t_{\mathbf{w}} \in \mathcal{T}} \mathbb{E}_\sigma \left[\prod_{i=1}^n \exp\{s (\sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}))\} \right] \quad (55)$$

$$= \sum_{t_{\mathbf{w}} \in \mathcal{T}} \prod_{i=1}^n \mathbb{E}_{\sigma_i} \left[\exp\{s (\sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}))\} \right] \quad (56)$$

$$\leq \sum_{t_{\mathbf{w}} \in \mathcal{T}} \prod_{i=1}^n \exp\left\{\frac{4s^2 \mathbf{W}_{uj}^2}{8}\right\} \quad (57)$$

$$= \sum_{t_{\mathbf{w}} \in \mathcal{T}} \exp\left\{\frac{4ns^2 \mathbf{W}_{uj}^2}{8}\right\} \quad (58)$$

$$\leq |\mathcal{T}| \sup_{t_{\mathbf{w}} \in \mathcal{T}} \exp\left\{\frac{4ns^2 \mathbf{W}_{uj}^2}{8}\right\} \quad (59)$$

$$= |\mathcal{T}| \exp\left\{\frac{ns^2 W^2}{2}\right\} \implies \quad (60)$$

$$\exp\{\mathbb{E}_\sigma \left[s \sup_{t_{\mathbf{w}} \in \mathcal{T}} \left(\sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right) \right]\} \leq \frac{\ln |\mathcal{T}|}{s} + \frac{nsW^2}{2} \implies \quad (61)$$

$$\exp\{\mathbb{E}_\sigma \left[s \sup_{t_{\mathbf{w}} \in \mathcal{T}} \left(\sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right) \right]\} \leq W \sqrt{2n \ln |\mathcal{T}|} \implies \quad (62)$$

$$\hat{\mathfrak{R}}_s(\mathcal{T}) = \mathbb{E}_\sigma \left[\sup_{t_{\mathbf{w}} \in \mathcal{T}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i t_{\mathbf{w}}(\mathbf{x}^{(i)}) \right) \right] \leq \frac{W \sqrt{2n \ln |\mathcal{T}|}}{n} \quad (63)$$

□

Inequality 52 uses Jensen's inequality, and equation 56 uses the independence property of expectation. To obtain 57, we first notice that $sgm() \in (0, 1)$, thus $t_{\mathbf{w}}(\mathbf{x}_i) \in (0, \mathbf{W}_{uj})$ and $\sigma_i t_{\mathbf{w}}(\mathbf{x}_i) \in (-\mathbf{W}_{uj}, \mathbf{W}_{uj})$, and then use Hoeffding's Inequality. Inequality 59 uses our assumption that $\|\mathbf{W}\|_{\max} \leq W$. By taking derivative of the RHS of 61 and set it to 0, we obtained $s = \sqrt{\frac{2 \ln |\mathcal{T}|}{nW}}$ hence we obtain equation 62. By dividing both sides by n we obtain equation 63.

Corollary 1. *Let $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^k\}$, $\mathcal{S} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ be a data set of n samples. Given a restricted Boltzmann machine with k visible units and m hidden ones, and it is trained by using*

CD-1 algorithm. For all the parameters $\theta = \{\mathbf{W}\}$, and $\mathbf{W} \in \mathbb{R}^{k \times m}$, assuming \mathbf{W} is bounded by spheres $\|\mathbf{W}\|_{\max} = \forall j \|\mathbf{W}_{\cdot j}\|_1 \leq W$, where $\mathbf{W}_{\cdot j}$ is the j -th column of \mathbf{W} . Let \mathcal{T} be a compositional function of \mathbf{x} with parameters \mathbf{W} , i.e.,

$$\mathcal{T} = \{\mathbf{W}_{ij} \text{sgm} \left(\sum_{v=1}^m \mathbf{W}_{iv} \text{sgm}(\mathbf{x}^T \mathbf{W}_{\cdot v}) \right) \mid \mathbf{W} \in \mathbb{R}^{k \times m} \forall j \in \{1, \dots, m\}\}. \quad (64)$$

We further assume the VC-dimension of \mathcal{T} is $VC(\mathcal{T})$. We can bound the empirical Rademacher complexity for the likelihood of this restricted Boltzmann machine as:

$$\hat{\mathfrak{R}}_s(\ln p_{\theta}) \leq \frac{W}{\sqrt{n}} \left(m\sqrt{2 \ln k} + k\sqrt{2VC(\mathcal{T}) \ln(n+1)} \right). \quad (65)$$

Proof. Using the result of **Remark 1**, we can bound $\hat{\mathfrak{R}}_s(\log Z_{\theta})$ in equation 44 by $\hat{\mathfrak{R}}_s(\log Z_{\theta}) \leq \hat{\mathfrak{R}}_s(\sum_{i=1}^k \mathcal{T}_i)$. Similar to equation 42, $\hat{\mathfrak{R}}_s(\sum_{i=1}^k \mathcal{T}_i) \leq \sum_{i=1}^k \hat{\mathfrak{R}}_s(\mathcal{T}_i)$. Knowing the fact that each \mathcal{T}_i is from the same hypothesis space further provides us with

$$\sum_{i=1}^k \hat{\mathfrak{R}}_s(\mathcal{T}_i) = k\hat{\mathfrak{R}}_s(\mathcal{T}). \quad (66)$$

Using the results in **Lemma 4** we obtain $\hat{\mathfrak{R}}_s(\log Z_{\theta}) \leq \frac{Wk\sqrt{2n \ln |\mathcal{T}|}}{n}$. Then by Sauer-Shelah lemma, we know

$$\max_{\mathcal{S}} |\mathcal{T}(\mathcal{S})| \leq (n+1)^{VC(\mathcal{T})}. \quad (67)$$

Therefore we obtain

$$\hat{\mathfrak{R}}_s(\log Z_{\theta}) \leq \frac{Wk\sqrt{2n \ln |\mathcal{T}|}}{n} \leq \frac{Wk\sqrt{2VC(\mathcal{T})n \ln(n+1)}}{n} \quad (68)$$

Together with the results from **Theorem 1**, while ignoring the bias term, we proved this corollary. \square

5 Future Direction

Can we get a tighter bound on it? Can we extend this results to multi-layer Boltzmann machines, like deep belief networks (DBN) or deep Boltzmann machines (DBM)? Is that possible to obtain the exact expression of the VC dimension of our constructed function \mathcal{T} ?

References

- Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009. ISSN 1935-8237.
- Geoffrey Hinton and Ruslan Salakhutdinov. A better way to pretrain deep Boltzmann machines. *Advances in Neural Information*, (3):1–9, 2012.

Geoffrey E. Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313(5786):504–7, 2006.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–54, 2006.

Jean Honorio. Lipschitz parametrization of probabilistic graphical models. *Uncertainty in Artificial Intelligence 2012*, 2012.

Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann Machines. *Artificial Intelligence*, 5(2):448–455, 2009.

Ruslan Salakhutdinov and Geoffrey Hinton. An Efficient Learning Procedure for Deep Boltzmann Machines. *Neural Computation*, 24(8):1967–2006, 2012.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, 1(1):194–281, 1986.